Outils de corpus

Corpus - 1

Clément Plancq

http://clement.plancq.github.io/outils-corpus

18 mars 2019

Introduction

Objectifs du cours

- Rappeler quelques notions clés sur les corpus
- Appliquer des commandes UNIX au traitement de corpus
- Utiliser des outils dédiés au traitement de corpus

Introduction (2)

A quoi servent les corpus ? : recherche

- linguistique
- TAL (modèles de langage)
- IA

Introduction (3)

A quoi servent les corpus ? (2) : industrie

- correction orthographique, grammaticale, stylistique (Cordial, MS Word, . . .)
- fouille de textes, acquisition de connaissances
- fouille d'opinions (e-reputation)
- extraction d'informations et systèmes de QA (Watson)
- traduction (Google Translate, DeepL, Skype translator) et aide à la traduction
- résumé automatique
- génération de texte
- reconnaissance vocale, synthèse vocale
- dialogue Homme-Machine, agents conversationnels (ELIZA, Siri (Apple), Alexa (Amazon), Cortana (Ms), Google Home)

Introduction (4)

Qu'est-ce qu'un corpus ?

A corpus in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a **finite-sized** body of **machine-readable** text, **sampled** in order to be maximally **representative of the language variety under consideration**. (McEnery and Wilson 2001)

Introduction (5)

Critères de définition d'un corpus

- Taille (en nb de mots le plus souvent)
- Annotations (niveaux et outils utilisés)
- Statut de la documentation (guidelines, publication scientifique)
- Stratégie d'échantillonage et origine des textes (genre)
- Modalité (écrit/oral)
- Licence et droits d'utilisation

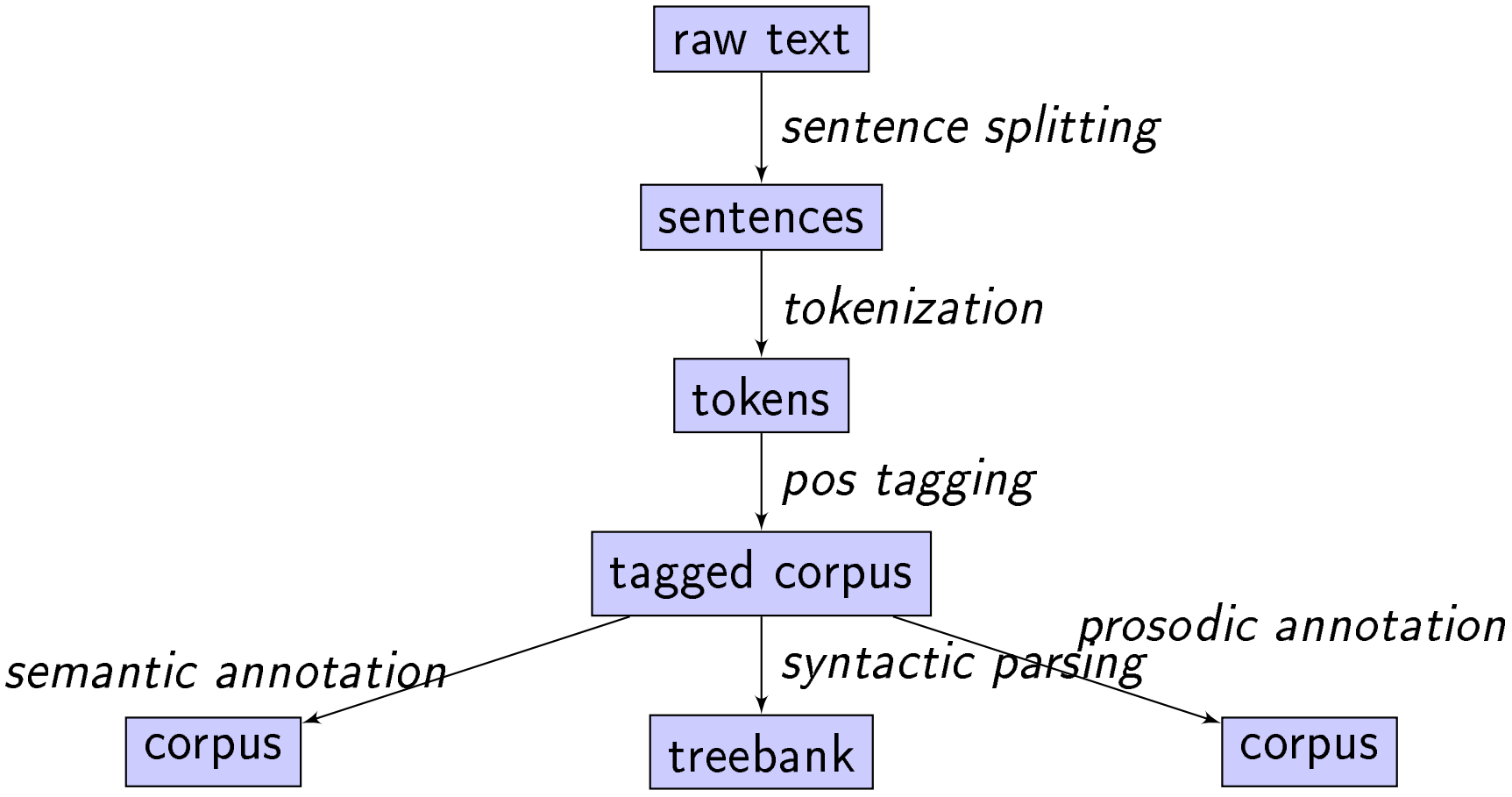
Introduction (6)

Types d'annotation

- (Transcription phonétique et/ou orthographique d'enregistrements)
- Annotation morpho-syntaxique (POS tagging)
- Annotation syntaxique (en constituants, en dépendance)
- Annotation sémantique
- Annotation en structures discursives

Making-of : from raw text to annotated corpora

Each step is error prone and depends on the accuracy of the preceding.



Commandes UNIX

Syntaxe de l'appel d'une commande UNIX

```
nom [-options] [argument 1...]
```

Exemple

```
ls -l kafka-metamorphosis_gutenberg.txt
```

man

La commande man permet d'accéder à la page de manuel d'une commande

```
man ls
```

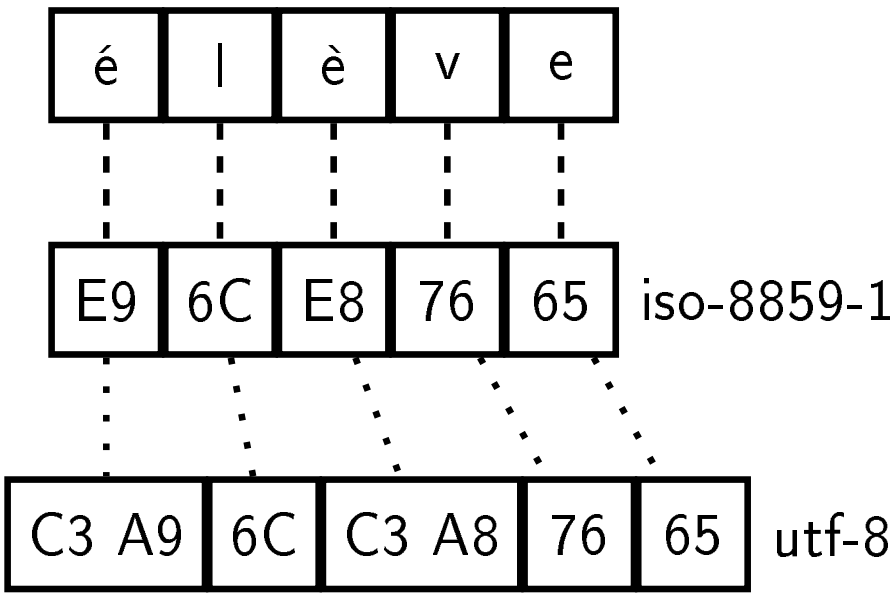
Codages de caractères

- Le codage de caractère en informatique associe une valeur numérique à un caractère
- Suivant le codage utilisé cette valeur numérique peut être représentée sur 7 bits (ASCII), 8 bits (ISO-8859-1, Windows-1252, MacRoman) ou un nombre d'octets variables (UTF-8)

`iconv`

- Créez un fichier nommé eleve.txt contenant le mot 'élève'
- `iconv -f utf-8 -t iso-8859-1 eleve.txt > eleve-latin1.txt`
- Comparez les tailles respectives des fichiers

Codages de caractères (2)



Compter les lignes et les mots

wc

`wc`	Compte les lignes, les mots et les octets d'un fichier
`wc -l`	Compte les lignes
`wc -w`	Compte les mots
`wc -b`	Compte les octets

Exercices

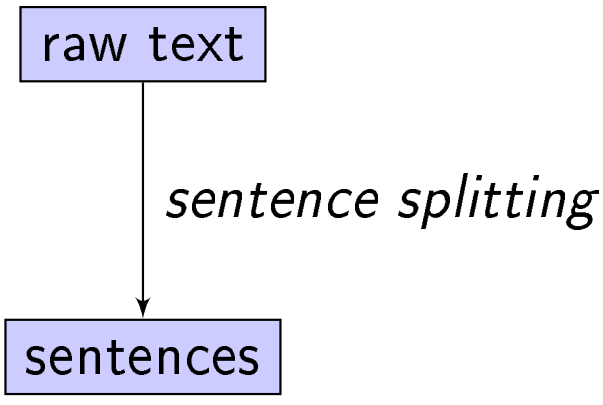
Comparez les lignes, mots et octets du fichier `kafka-metamorphosis_gutenberg.txt`

Compter les lignes et les mots (2)

Mise en garde

- Les lignes ne sont pas des phrases
- Les *mots graphiques* de `wc` ne correspondent pas forcément aux unités linguistiques *mots*

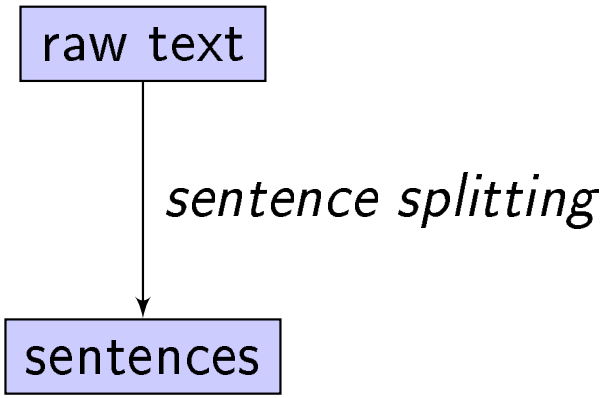
Sentence splitting



One morning, when Gregor Samsa woke from troubled dreams, he found himself transformed in his bed into a horrible vermin. He lay on his armour-like back, and if he lifted his head a little he could see his brown belly, slightly domed and divided by arches into stiff sections.

<s>One morning, when Gregor Samsa woke from troubled dreams, he found himself transformed in his bed into a horrible vermin.</s> <s> He lay on his armour-like back, and if he lifted his head a little he could see his brown belly, slightly domed and divided by arches into stiff sections.</s>

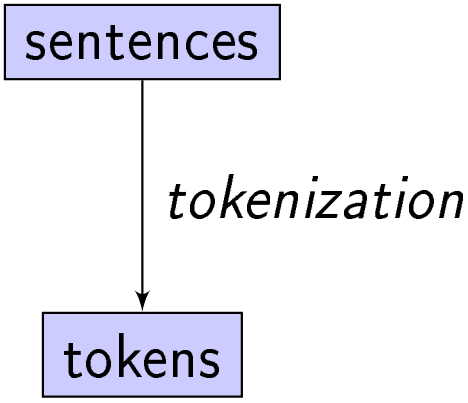
Sentence splitting



"Oh, God!" called his mother, who was already in tears, "he could be seriously ill and we're making him suffer. Grete! Grete!" she then cried. "Mother?" his sister called from the other side. They communicated across Gregor's room. "You'll have to go for the doctor straight away. Gregor is ill. Quick, get the doctor. Did you hear the way Gregor spoke just now?" "That was the voice of an animal", said the chief clerk, with a calmness that was in contrast with his mother's screams.

?

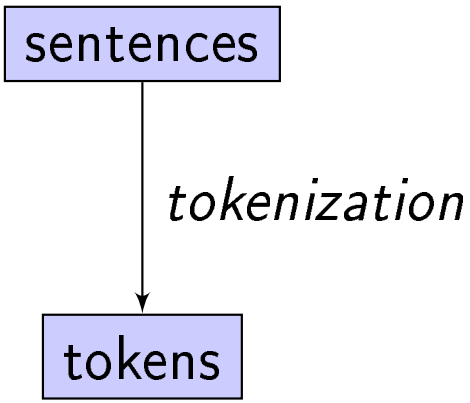
Tokenization




```
One morning, when Gregor Samsa woke from
troubled dreams, he found himself
transformed in his bed into a horrible
vermin.
```

```
One
morning,
when
Gregor
Samsa
woke
from
troubled
dreams,
he
found
himself
transformed
in
his
bed
into
a
horrible
vermin.
```

Tokenization




```
"What's happened to me?" he thought. It
wasn't a dream.
```

?

`wasn't` vs. `was not`

Tokenization has an impact on frequency count

Découpage en mots : tokenisation

tr, perl -pe

`tr 'e' 'i'` remplace tous les car. 'e' par des car. 'i'

`perl -pe 's/e/i/'` remplace tous les car. 'e' par des car. 'i'

exercice

Utilisez `tr` ou `perl -e 's///'` pour remplacer les espaces par des '\n' dans `kafka-metamorphosis_gutenberg.txt`.
Comptez les mots et comparez avec `wc -w`.
Comparez avec le découpage en mots de
`kafka-metamorphosis_gutenberg_treetagger.txt`

Tri et comptage

`sort`, `uniq -c`

`sort` tri par ordre alphanumérique

`uniq -c` supprime les lignes doublons et `-c` compte le nb de lignes identiques

exercice

Utilisez `tr`, `sort` et `uniq -c` dans un chaîne de traitement pour classer par ordre de fréquence décroissante les mots de `kafka-metamorphosis_gutenberg.txt`.